Supplementary Information Document

Phylogenetic conflict in bears identified by automated genome-wide discovery of transposable element insertions

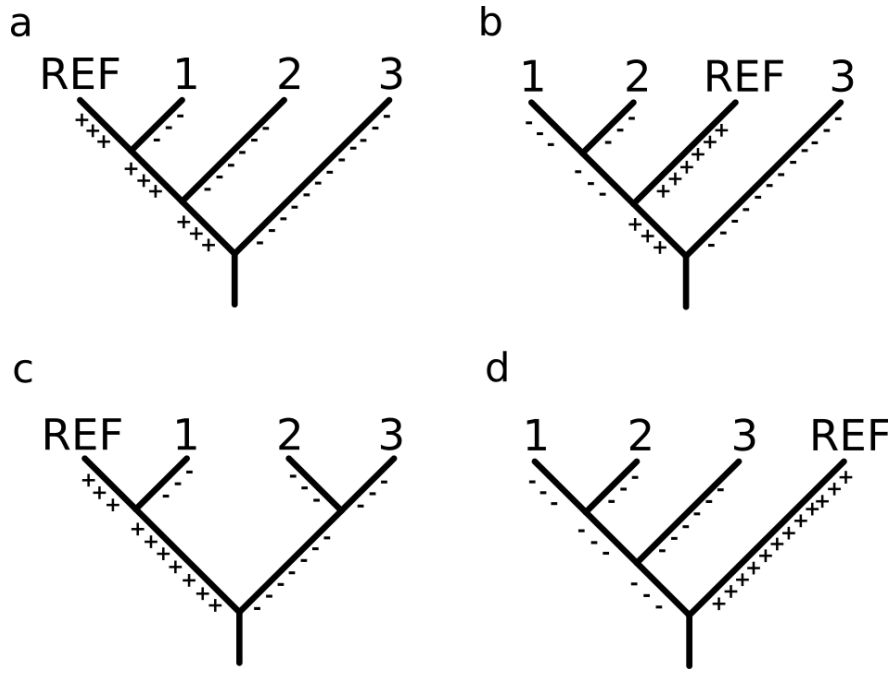
Fritjof Lammers[1,2], Susanne Gallus[1], Axel Janke[1,2], Maria A Nilsson[1§]

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany.
[2]Goethe University Frankfurt, Institute for Ecology, Evolution & Diversity, Biologicum, Max-von-Laue-Str.13, 60439 Frankfurt am Main, Germany.

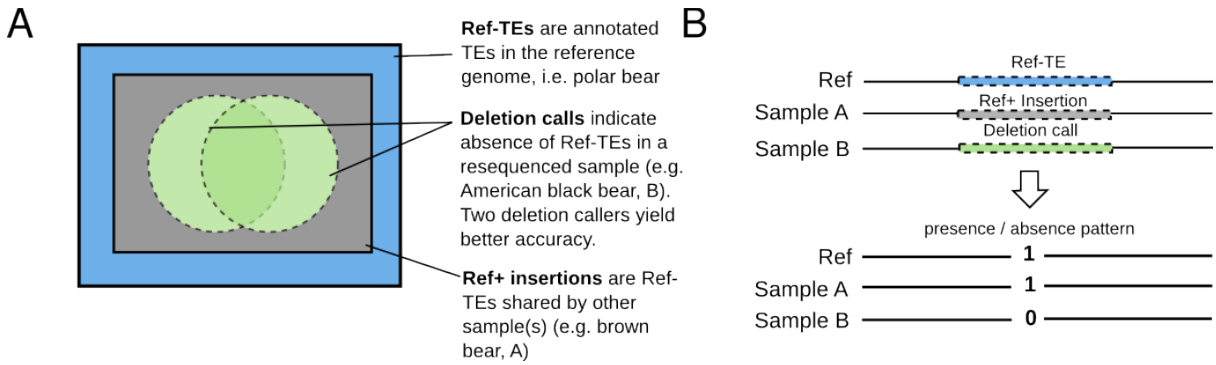
Supplementary Figures

Supplementary Tables

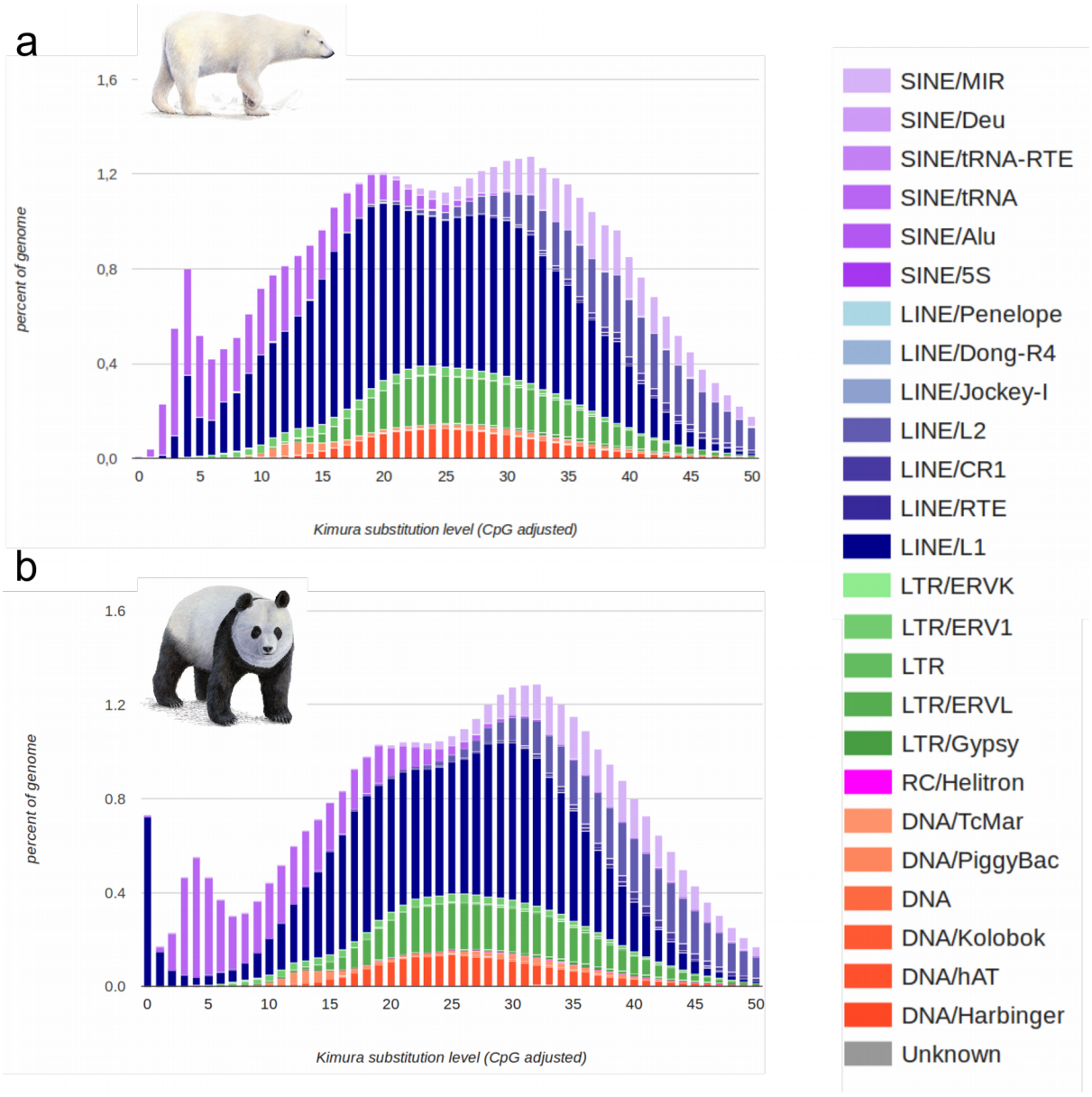
.



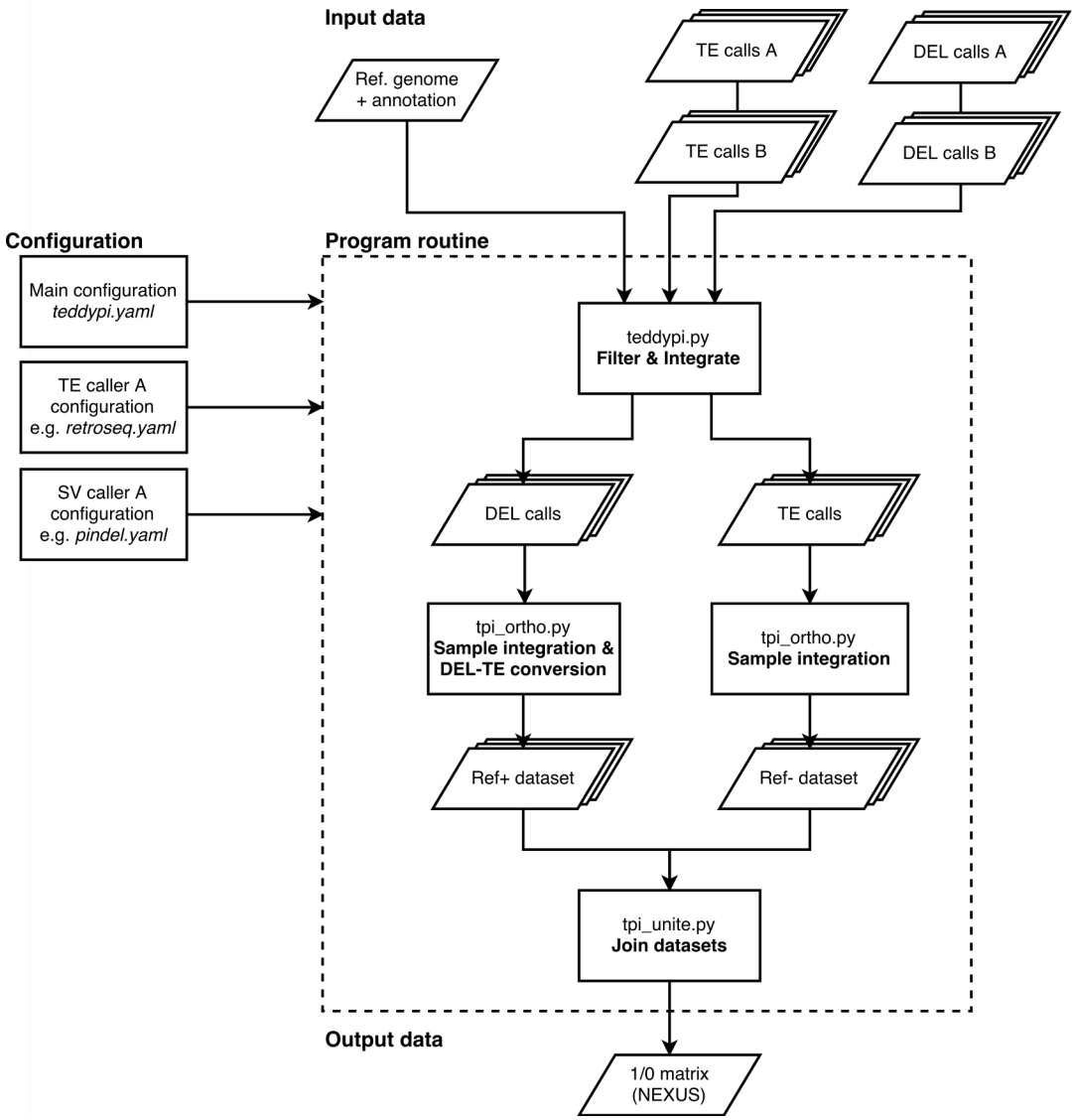
Supplementary Figure 1. Considerations for the choice of TE callers. Depending on the position of the reference genome (REF) within the expected species tree, insertions on different branches can be detected by non-reference (Ref-) insertion callers and/or by reference insertion-callers. In the different topologies A), B), C) the reference genome is nested inside the tree and not an outgroup. Ref- insertions (depicted as '-' at the branches) can therefore be detected on branches, that do not lead to the reference genome. The reference (Ref+) insertions (depicted as '+' along the branch) shared with additional taxa require Ref+-insertion calling to support the internode branches. D) In this example the reference genome is placed as the outgroup. Only in this case the sole use of a non-reference caller can find insertions supporting terminal and internode branches to taxon 1, 2 and 3. Depending on the initial phylogenetic hypothesis or knowledge about the taxa under study, a selection for Ref-, Ref+ or a combined detection approach needs to made.



Supplementary Figure 2. The principle of deletion calling. A) TEs annotated in the reference genome (Ref-TEs, blue area) can either be private to the reference genome or shared with other species, if the insertion occurred in a common ancestor of both. TEs shared by the reference and other species are thus a subset of all Ref-TEs (grey area). To infer presence/absence patterns for such TEs, sample genomes are screened for deletion signatures with two programs (green circles). The presence of a deletion signature thereby indicates the absence of the TE insertion. Combining two sets of deletions called by two programs (here Pindel and Breakdancer) minimizes recognizing false positive Ref+ calls. B) A schematic alignment shows how the information of annotated reference TEs and the presence of a deletion call in sample B indicates a Ref+ insertion in sample A. If sequencing reads map normally and no deletion or any other structural variant is discovered by a TE or SV caller, the presence of the same TE in sample A and the reference assembly is inferred. For sample B at least one deletion call was made. Subsequently, the locus is recorded as 1-1-0 for the presence of TE insertions.

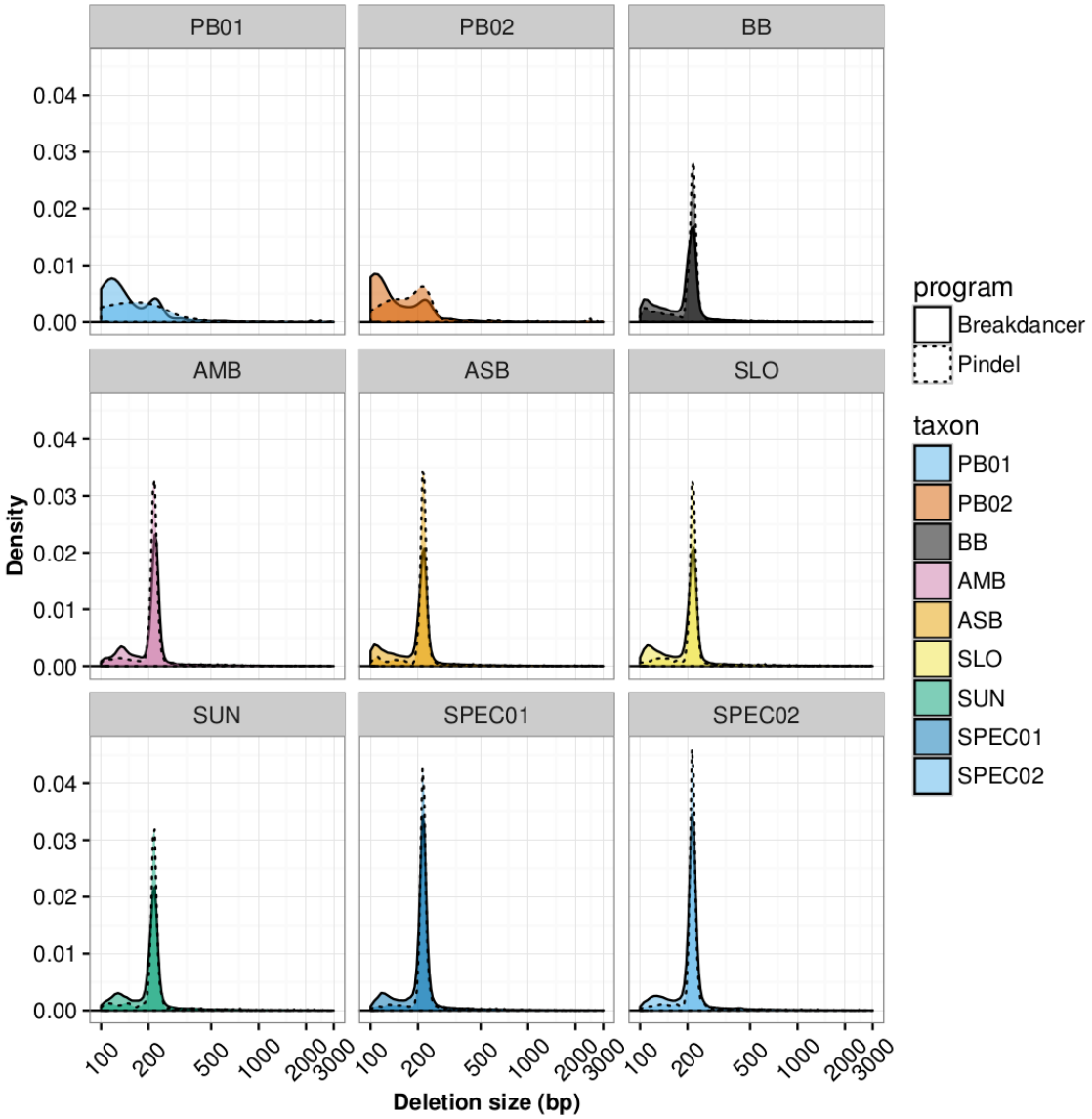


Supplementary Figure 3. Repeat landscapes from genome assemblies of polar bear (A) and giant panda (B). The graphs show the relative amount of each transposable element group in the genome in bins of 1% divergence to their consensus sequences. The divergence is shown on the x-axis and calculated as CpG adjusted Kimura-2-parameter substitutions to the consensus sequences. The y-axis shows the percentage of genome coverage for each TE. The repeat landscape for polar bear was generated with RepeatMasker. The repeat landscape for giant panda was copied from http://repeatmasker.org/species/ailMel.html [last accessed 2016/05/03].

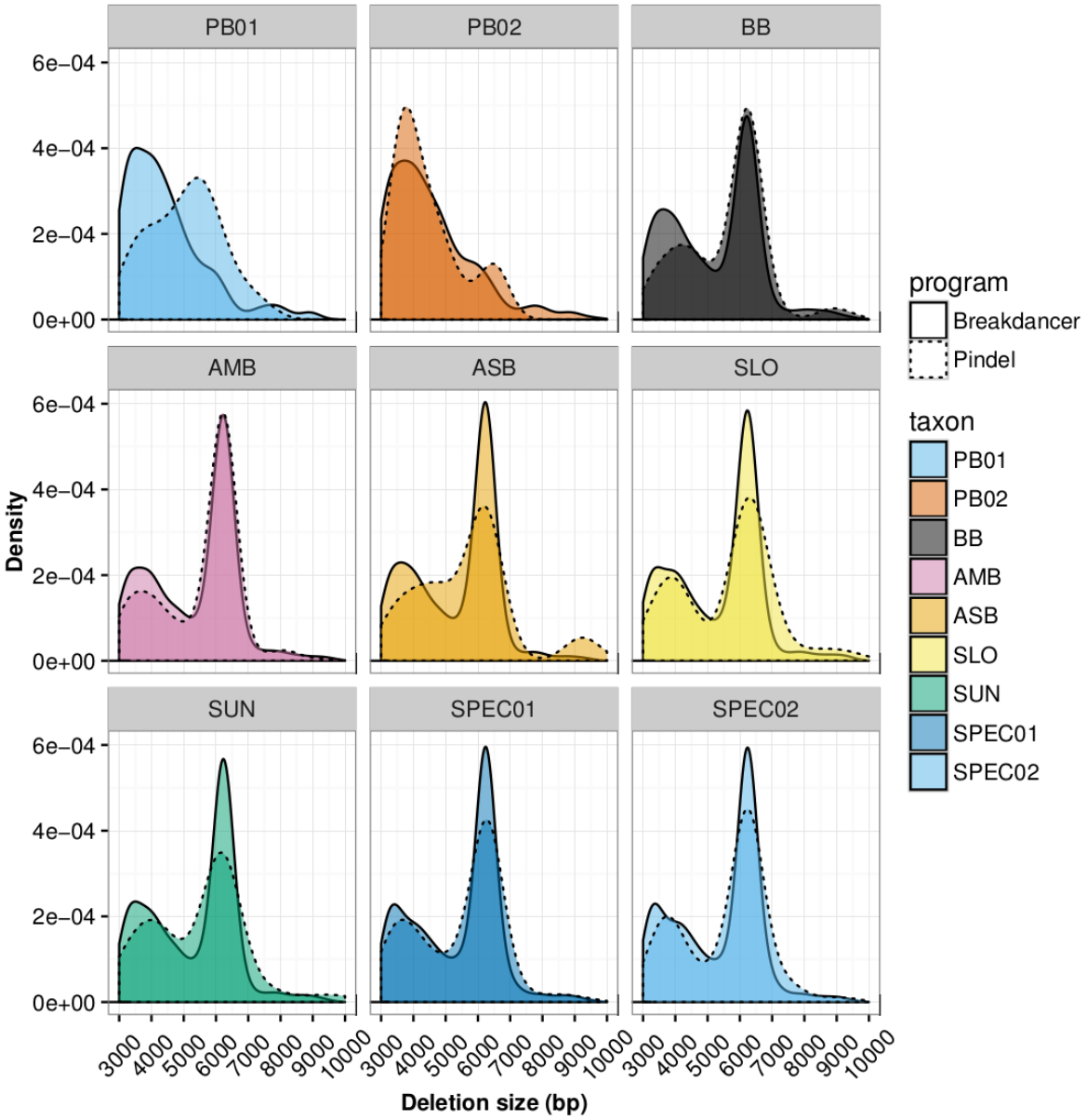


Supplementary Figure 4. Flowchart of the TeddyPi pipeline. Input data: TeddyPi requires a reference genome (Ref. Genome, FASTA format) and annotation of repetitive regions and assembly gaps (as BED files). TE and SV calls from resequenced samples, that were mapped against the reference genome are processed and data from multiple TE/SV callers (denoted as A and B) can be utilized (VCF files). TeddyPi is configured with a main configuration file that stores parameters and information on samples. Additional configuration files for each utilized TE/SV caller are needed and define the filtering steps applied to the data. The configuration files are read by all modules in the program routine of TeddyPi.

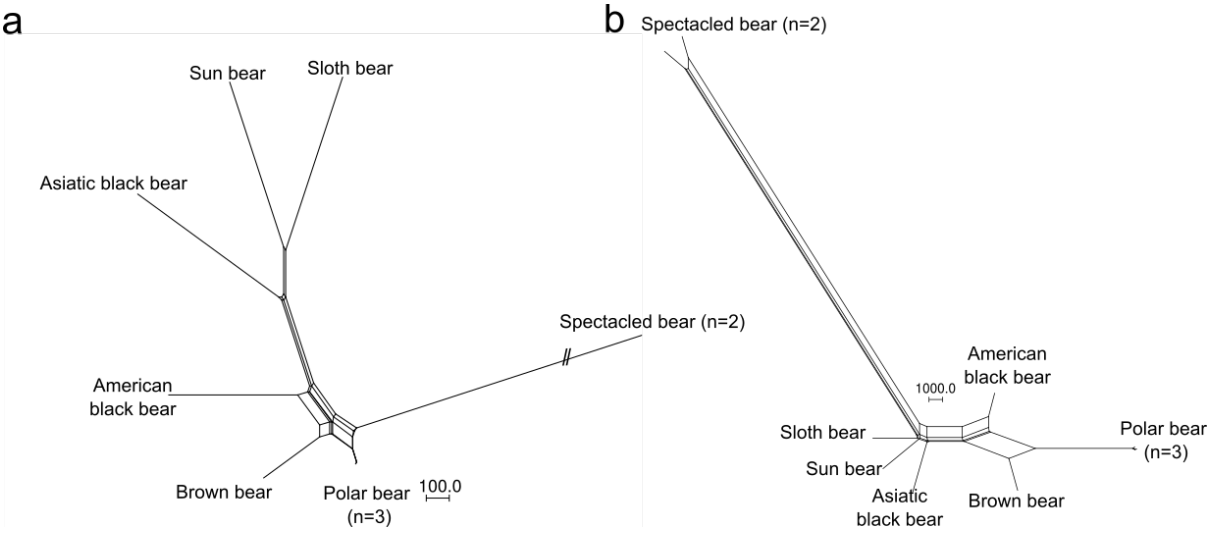
In the routine, first TE/SV callsets are filtered and if more than one are supplied all TE and all deletion (DEL) calls are integrated to a non-redundant set or by intersection (teddypi.py). The processing and filter methods are defined in tpi_filter.py (not shown). For all samples TE and DEL datasets are given as intermediate output. In tpi_ortho.py, data from all samples are integrated for TE and DEL calls respectively to create the Ref+ and Ref- datasets. Finally, both datasets are unified in tpi_unite.py and a presence/absence is stored in NEXUS format.



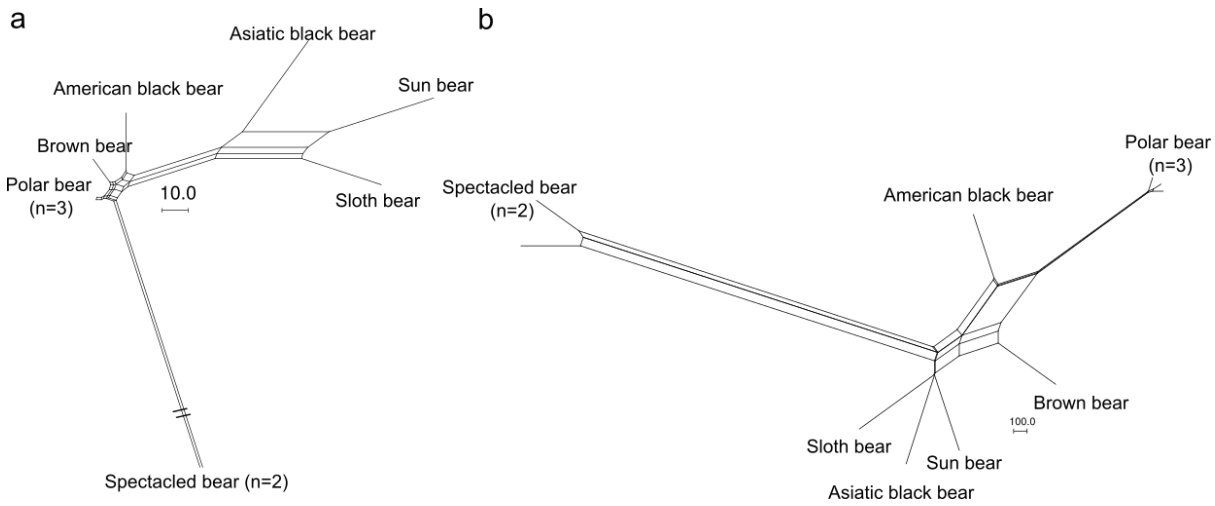
Supplementary Figure 5. Length distribution for deletions called by Pindel (dotted lines) and Breakdancer (solid lines) for each analyzed genome. The plots show the density for deletion of lengths between 100 and 3,000 bp. The x-axis is log10 scaled. A peak around 200 bp indicates deletions originated by SINE insertions. Sample names are polar bear (PB), brown bear (BB), American black bear (AMB), Asiatic black bear (ASB), sloth bear (SLO), sun bear (SUN), and spectacled bear (SPEC).



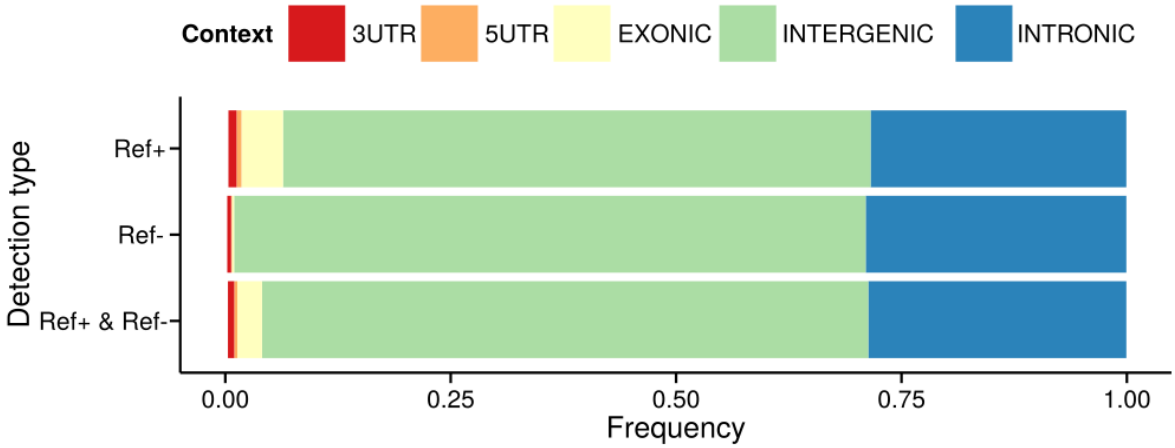
Supplementary Figure 6. Length distribution for deletions called by Pindel (dotted lines) and Breakdancer (solid lines) for each analyzed genome. The plots show the density for deletion of lengths between 3,000 bp and 10,000 kb. A peak around 6,000 bp indicates deletions originated by full length LINE-1 insertions. Shorter peaks likely originate from 5'truncated LINEs. The relative abundance of the 6,000 bp peak is similar in all samples, except polar bear, which exhibits a greater extent of deletion <6,000 bp. Sample names are polar bear (PB), brown bear (BB), American black bear (AMB), Asian black bear (ASB), sloth bear (SLO), sun bear (SUN), and spectacled bear (SPEC).



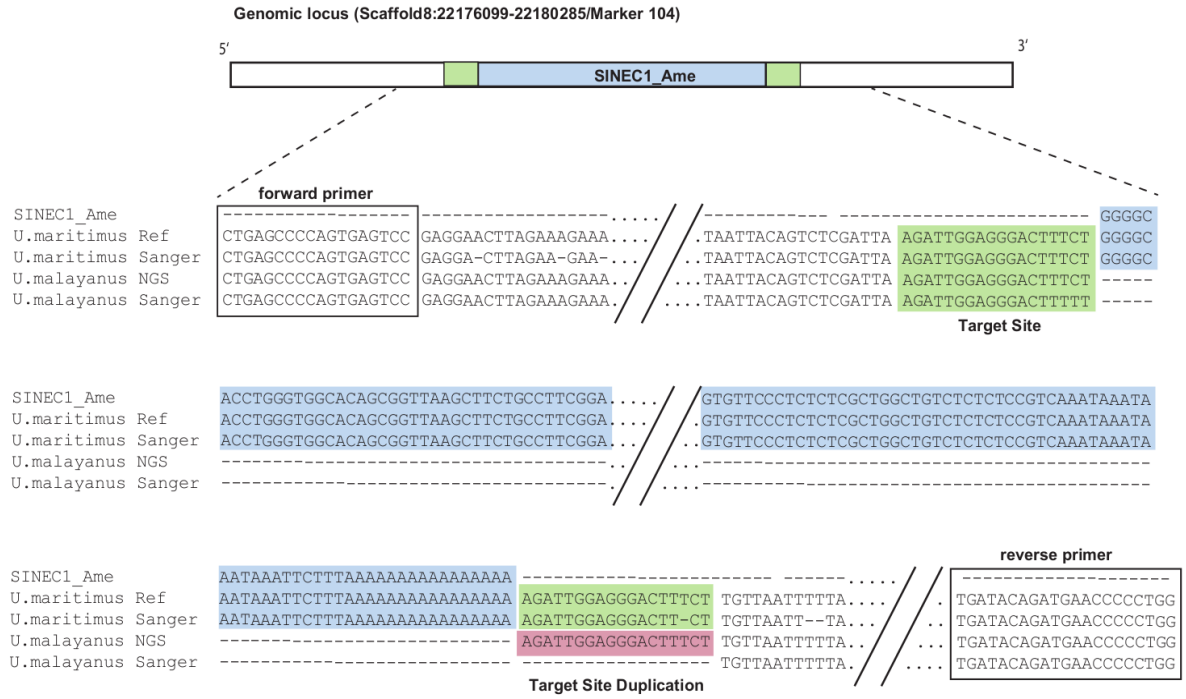
Supplementary Figure 7. Phylogenetic networks reconstructed from SINE insertions shown separately for Ref- insertions (A) and Ref+ insertions (B). A) Parsimony splits network from 61,026 Ref- SINE insertions have better resolution for the relationship between Asiatic black, sun and sloth bear than among polar bear, American black and brown bear. B) Parsimony splits network from 71,067 Ref+ SINE insertions resolve the relationship among polar bear, brown bear and American black bear. The edges between the three Asiatic bears are short and allow only limited resolution, but are consistent with the species tree.



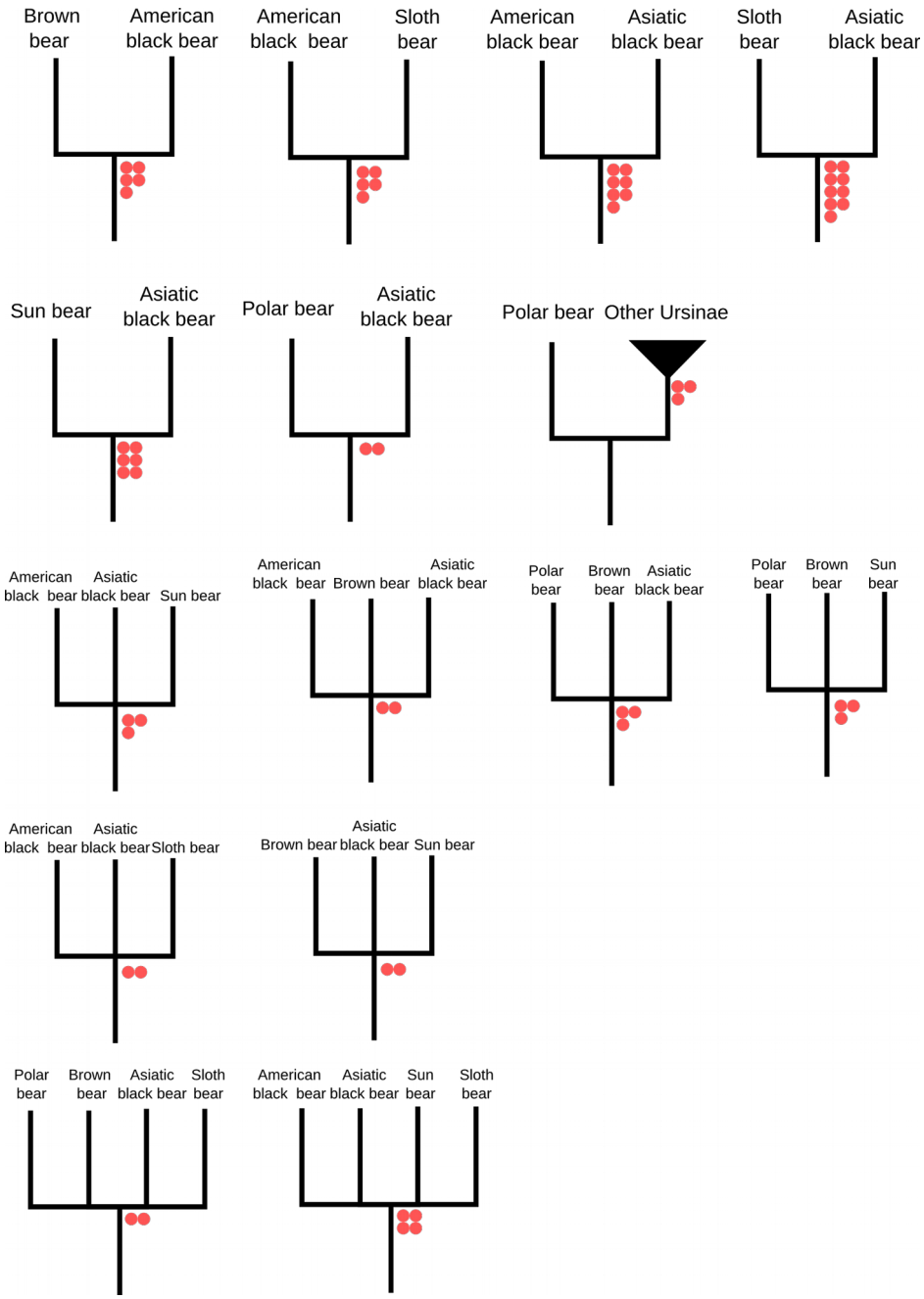
Supplementary Figure 8. Phylogenetic networks reconstructed from LINE1 insertions shown separately for Ref- insertions (A) and Ref+ insertions (B). A) A parsimony splits network from 6,455 LINE1 Ref- insertions with a minimum threshold of one character per branch. As for SINEs (Fig S5), Ref- insertions have better resolution for the relationship between Asiatic black, sun and sloth bear than for polar bear, brown bear and American black bear, respectively. The nested position of the reference genome used for TE calling causes the polar bear to appear in the center of the network. B) Parsimony splits network calculated from 11,965 LINE1 Ref+ insertions (threshold 1 character per edge) produce a long edge to the polar bear, brown bear plus American black bear, supporting a sister group relationship between them. The edges between the Asiatic bears are short and show only limited resolution from this type of marker.



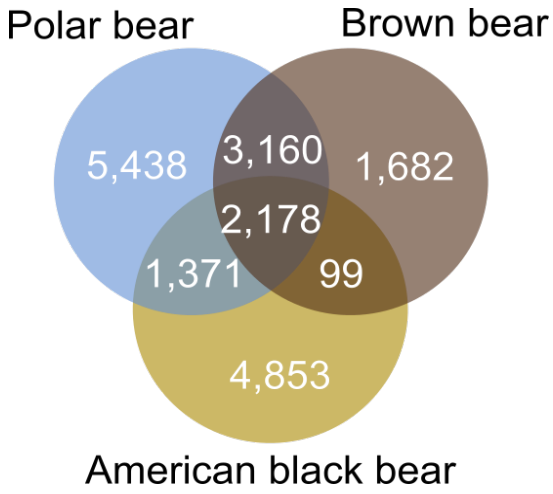
Supplementary Figure 9. Insertion frequency of TEs (SINEs and LINEs combined) into different genomic contexts in the polar bear genome. Color coding for genomic contexts is explained in the legend; the frequency describe the relative amount of TE insertions found in the respective genomic background. The insertion frequency is given separately for the different detection types: reference insertions (Ref+), non-reference insertion (Ref-) and the combined dataset. As expected, most insertions occurred in non-coding regions, i.e. intergenic regions and introns.



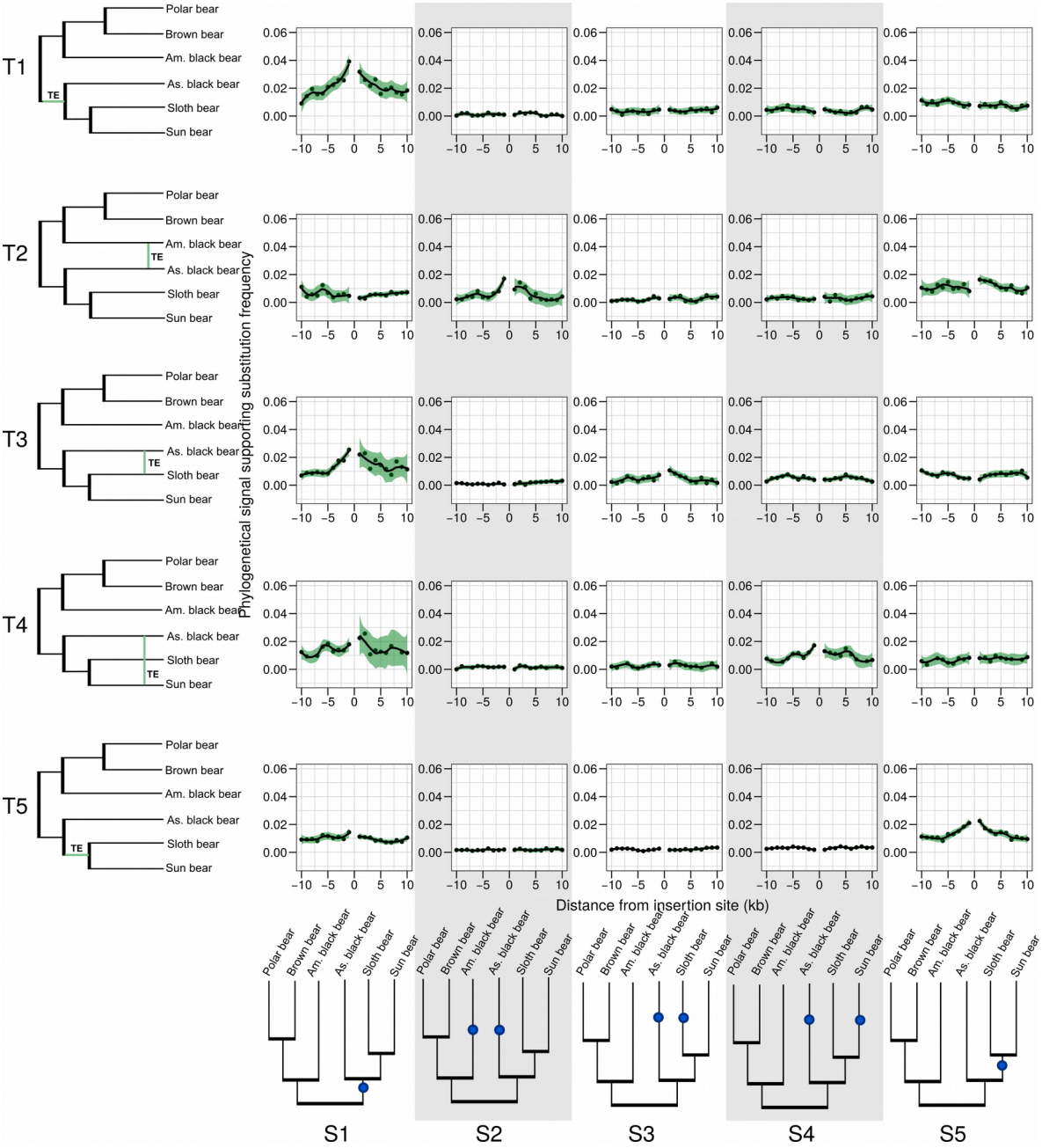
Supplementary Figure 10. Alignment of marker 104. Alignment of a genomic locus that harbors a reference SINEC1_Ame insertion in the polar bear (*U. maritimus*) while the insertion is absent in the sun bear (*U. malayanus*). The target site duplication (TSD) are highlighted and positions of conserved primers are boxed. For *U. maritimus* the genome assembly sequence (Ref) and the Sanger validated sequence are shown (Sanger), for *U. malayanus* the Illumina-based consensus (NGS) and Sanger sequences are shown. Note that, the consensus sequence in the Illumina sequence have a false target site duplication (TSD, highlighted in red). Sanger sequencing of the same sample used for Illumina sequencing revealed the absence of the SINE insertion and the TSD in *U. malayanus.*



Supplementary Figure 11. Phylogenetic signal from TE markers that are species-tree incongruent based on validation experiments. The topologies include phylogenetic signals that match phylogenetic hypotheses selected from the *in silico* predictions (e.g. shared TE insertions for Asiatic black and sloth bear, as shown in Table S12) and signals that contradict the *in silico* predictions, i.e. markers with partially erroneous predictions. Signals with only one supporting TE insertions are not shown graphically and listed in Data S1. TE insertions found in more than two species, are drawn on polytomic trees. Each red dot represents one TE insertion.



Supplementary Figure 12. Venn Diagram showing conflict among Polar bear, brown bear and American black bear on basis of inferred SINE insertions. The insertions numbers were extracted from the presence/absence table.



Supplementary Figure 14. Phylogenetic signals in the genomic sequences flanking the TEs. The panels show the frequencies of substitutions that support specific phylogenetic signals (specified in columns S1-S5) in windows of 1 kb surrounding the TE (from -10 to 10 kb) insertion site among loci carrying TE insertions for different phylogenetic signals (rows T1-T5). Each row (T1-T5) represents TE loci as indicated by the phylogenetic tree on the left-hand side (green branches). Vertical bars in trees T2-T4 connect branches in which the TE is present; they indicate TE insertions representing a phylogenetic signal incongruent to the species tree. The columns S1-S5 indicate substitutions with different phylogenetic signals. The signal is shown as blue dot in the trees in the bottom. For example, the first panel (T1-S1), show the frequency of substitutions that are shared by Asiatic black, sun and sloth bear (S1) in TE loci that show the same phylogenetic signal (T1). The second panel (T1-S2), shows the frequency of substitutions shared by American and Asiatic black bear in the same set of TE loci, as before (both are in row T1). For a detailed discussion refer to Supplementary Note 2.

Supplementary Table 1. List of genomes analyzed in this study.

Binomial name	Common name	ID	Accession number	Coverage	Insert size
Ursus maritimus	Polar bear[1]	PBREF	http://gigadb.org/dataset/100008	100X	
Ursus maritimus	Polar bear	PB01	SRR518686, SRR518687	11.8X	241 bp, SD= 19.8
Ursus maritimus	Polar bear	PB02	SRR518661, SRR518662	12.15X	267 bp, SD=31.3
Ursus arctos	Brown bear	BB	SRR935592, SRR935595, SRR935624, SRR935628	18.97X1	479 bp, SD= 22.6
Ursus americanus	American black bear	AMB	SRR518723	19.31X	300 bp, SD= 46.2
Ursus malayanus	Malayan Sun Bear	SUN	PRJEB9724	9.05	471 bp, SD= 32.4
Ursus ursinus	Sloth bear	SLO	PRJEB9724	9.09X	482 bp, SD= 23.5
Ursus thibetanus	Asiatic black bear	ASB	PRJEB9724	9.92X	482 bp, SD= 27.8
Tremarctos ornatus	Spectacled bear	SPE01	PRJEB9724	9.62X	476 bp, SD= 32.1
Tremarctos ornatus	Spectacled bear	SPE02	PRJEB9724	9.37X	474 bp, SD= 40.3

Note - [1]denotes the reference sequence, named polar bear genome; [2]SD = standard deviation

Supplementary Table 2. Selected phylogenetic hypotheses subject to validation experiments. The table shows the different taxon sets, for which synapomorphic TE loci were selected from the *in silico* data set. For each set, the range of primer IDs is given as well as a brief description for each hypothesis' origin. Hypotheses were based on the species tree or alternative phylogenetic trees proposed for Ursidae. For primers 120-179, loci were selected randomly for phylogenetically informative TE insertion from the specified datasets. Primer 1-29, were used to preliminary experiments and therefore not listed.

Primer ID	TE presence / Selection criterion	Description
30-39	Asiatic black bear, sun bear, and sloth bear	"Species tree"
40-49	Sloth bear and sun bear	"Species tree"
50-59	Sloth bear and American black bear	Autosomal / Y genes (Kutschera et al. 2014)
60-69	American black bear, Asiatic black bear, and sun bear	mtDNA tree
70-79	American black bear and Asiatic black bear	mtDNA tree
80-89	Sun bear and Asiatic black bear	Alternative topology from autosomal genes
90-99	Brown bear, American black bear, and polar bear	"Species tree"
100-119	Polar bear, brown bear, Asiatic black bear, and sun bear	
120-139	random Ref+	
140-149	random RetroSeq	
150-159	random Pindel	
160-169	random Breakdancer	
170-179	random RetroSeq + Mobster	

Supplementary Table 3. Repetitive elements in the polar bear genome sequence. The genome has been screened with RepeatMasker for Carnivore specific repeats in strict mode.

Element	Number of elements	Length (bp)	% of genome
SINEs	1,223,168	194,257,084	8.42
Alu/B1	0	0	0
MIRs	499,702	74,868,620	3.24
LINEs	978,888	492,525,513	21.34
LINE1	561,205	380,530,435	16.48
LINE2	350,263	96,987,569	4.20
L3/CR1	48,769	10,829,082	0.47
RTE	16,999	3,949,655	0.17
LTR	320,346	122,027,132	5.29
ERVL	92,885	40,217,869	1.74
ERVL-MaLRs	150,576	52,141,355	2.26
ERV_classI	50,572	22,967,428	0.99
ERV_classII	1,105	510,735	0.02
DNA	340,447	70,582,462	3.06
hAT-Charlie	196,435	37,647,590	1.63
TcMar-Tigger	50,119	14,904,611	0.65
Unclassified	6,539	1,107,006	0.05
Total interspersed repeats		880,499,197	38.14
Small RNA	751,454	121,127,388	5.25
Satellites	145	20,875	0
Simple repeats	632,864	28,020,132	1.21
Low complexity	103,265	5,288,810	0.23
Unmasked sequence			60.4

Supplementary Table 4. Prediction counts from RetroSeq SINE calls for raw calls and each filtering step. First, SINE insertion were selected from the dataset (SINEs). The call sets were filtered step-wise for homozygousity and quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage).

Sample	Raw	SINEs	Quality	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	10,007	4,487	208	174	92	91
Polar bear 02	11,929	4,831	301	233	150	149
Brown bear	10,587	20,655	4,634	4,135	3,059	3,056
American black bear	59,728	38,193	11,417	10,750	7,374	7,372
Asiatic black bear	44,597	28,520	9,129	8,540	6,731	6,725
Sloth bear	45,223	28,886	13,871	13,196	10,705	10,697
Sun bear	48,792	32,458	12,875	12,285	10,015	10,005
Spectacled bear 01	95,630	72,520	36,914	36,205	29,654	29,638
Spectacled bear 02	95,878	73,103	36,659	35,950	29,433	29,422
Total	696,041	303,653	126,008	121,468	97,213	97,155

Supplementary Table 5. Predictions counts from Mobster for raw calls and each filtering step for SINEs. First, SINE insertions with at least 4 supporting reads on 5' and 3' end of the insertion were selected from the dataset (column SINEs). The call sets were filtered step-wise for quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage). Note that Mobster does not give zygosity information.

Sample	Raw	SINEs	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	14,589	6,887	5,351	1,215	1,002
Polar bear 02	15,752	6,832	5,276	1,326	1,087
Brown bear	35,471	20,063	18,566	9,360	9,186
American black bear	73,865	47,542	44,857	14,701	14,439
Asiatic black bear	40,425	26,170	24,820	14,106	13,887
Sloth bear	41,772	28,074	26,672	15,530	15,362
Sun bear	47,494	32,905	31,432	17,117	16,833
Spectacled bear 01	84,380	66,838	65,297	40,466	40,221
Spectacled bear 02	88,578	70,278	68,652	40,725	40,509
Total	491,193	305,589	290,923	154,546	152,526

Supplementary Table 6. Prediction counts from RetroSeq LINE1 calls for raw calls and each filtering step. LINE1 insertions were selected from the dataset (L1s). The call sets were filtered step-wise for homozygosity and quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage).

Sample	LINE1	Quality	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	5,271	197	114	37	36
Polar bear 02	6,814	336	170	58	56
Brown bear	18,446	3,859	2,088	1,004	1,000
American black bear	20,004	5,168	4,058	2,330	2,327
Asiatic black bear	15,117	5,069	3,318	1,919	1,912
Sloth bear	15,251	5,854	4,253	2,669	2,663
Sun bear	15,310	5,580	4,082	2,634	2,630
Spectacled bear 01	21,288	10,013	9,145	6,186	6,174
Spectacled bear 02	20,965	9,750	8,866	5,972	5,964
Total	138,466	45,826	36,094	22,809	22,762

Supplementary Table 7. Predictions counts from Mobster for LINE1 calls and each filtering step. First, SINE insertions with at least 4 supporting reads on 5' and 3' end of the insertion were selected from the dataset (column SINEs). The call sets were filtered step-wise for quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage). Note that Mobster does not differentiate between homo- and heterozygosity.

Sample	LINE1	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	7,609	3,905	211	170
Polar bear 02	8,818	4,427	210	169
Brown bear	15,096	9,790	1,406	1,319
American black bear	25,394	20,003	1,751	1,691
Asiatic black bear	14,007	10,032	1,961	1,865
Sloth bear	13,504	9,903	2,053	1,989
Sun bear	14,338	10,628	2,303	2,213
Spectacled bear 01	17,184	15,029	4,382	4,256
Spectacled bear 02	17,892	15,683	4,415	4,297
Total	133,842	99,400	18,692	17,969

Supplementary Table 8. Summary of non-reference TE insertion counts in Ursinae for SINEs and LINEs with values from RetroSeq and Mobster and their overlap.

	SINEs			LINE1s		
Sample	**RetroSeq**	**Mobster**	**Overlap**	**RetroSeq**	**Mobster**	**Overlap**
Polar bear 01	91	1,002	65	36	170	8
Polar bear 02	149	1,087	120	56	169	14
Brown bear	3,056	9,186	2,518	1,000	1,319	221
American black bear	7,372	14,439	6,711	2,327	1,691	556
Asiatic black bear	6,725	13,887	5,727	1,912	1,865	729
Sloth bear	10,697	15,362	9,434	2,663	1,989	1,104
Sun bear	10,005	16,833	8,594	2,630	2,213	1,080
Spectacled bear 01	29,638	40,221	25,960	6,174	4,256	1,993
Spectacled bear 02	29,422	40,509	25,330	5,964	4,297	2,029
Total	97,155	152,526	84,462	22,762	17,969	7,734

Supplementary Table 9. Filtering results for the Breakdancer dataset. From the raw dataset, deletions were selected and filtered for a length >100 bp and < 10 kb (Size). Then, the call sets were filtered step-wise for vicinity to assembly gaps (Filtered Gaps) and for vicinity or overlaps with satellite DNA, and other repetitive sequences in polar bear reference genome that were not an interspersed repeat (Repeat-filtered). Finally, call sets were filtered for regions of extraordinary high coverage (Coverage).

				Filtered		Coverag
Sample	**Raw**	**Deletion**	**Size**	**Gaps**	**Repeat-filtered**	**e**
Polar bear 01	5,079	3,963	2,383	1,702	1,403	1,337
Polar bear 02	4,675	3,791	3,348	2,520	2,133	2,033
Brown bear	57,097	35,210	27,088	24,576	23,082	22,986
Am black bear	33,191	29,303	29,036	26,756	23,893	23,607
As black bear	69,487	43,393	38,800	35,939	33,620	33,465
Sloth bear	72,885	44,724	40,811	37,829	35,212	35,040
Sun bear	70,014	43,155	39,109	36,284	33,871	33,638
Spectacled bear 01	146,473	87,980	80,994	77,021	72,405	72,127
Spectacled bear 02	143,694	92,061	83,477	79,347	72,689	71,780
total	602,595	383,580	345,046	321,974	298,308	296,013

Supplementary Table 10. Filtering results for the Pindel dataset. From the raw dataset, deletions, homozygous deletions were selected and filtered for a length >100 bp and < 10 kb (Size). Then, the call sets were filtered step-wise for vicinity to assembly gaps (Filtered Gaps) and for vicinity or overlaps with satellite DNA, and other repetitive sequences in polar bear reference genome that were not an interspersed repeat (Repeat-filtered). Finally, call sets were filtered for regions of extraordinary high coverage (Coverage).

Sample	Raw	Deletion	Size	Filtered Gaps	Repeats-filter	Coverage
Polar bear 01	155,489	8,842	111	75	54	43
Polar bear 02	157,895	8,932	124	80	64	54
Brown bear	737,502	84,769	2,561	2,311	1,910	1,863
Am black bear	1,031,231	204,739	5,503	5,145	4,312	4,092
As black bear	758,266	74,660	1,213	1,103	935	883
Sloth bear	829,500	96,421	1,159	1,054	877	805
Sun bear	785,786	87,122	1,060	970	805	733
Spectacled bear 01	1,357,194	236,097	3,207	3,027	2,603	2,344
Spectacled bear 02	1,320,685	213,305	2,776	2,613	2,273	2,048
Total	11,746,167	1,014,887	17,714	16,378	13,833	12,865

Supplementary Table 11. Results of Ref+ insertion processing. Deletion calls from Breakdancer and Pindel were combined to a non-redundant set (DEL_nr) and screened for intersection with TEs in the polar bear reference genome (TEs), from which calls corresponding to SINE and LINE1 insertions were extracted (SINEs, LINE1s). Other deletions corresponding to TE insertions are counted as 'Other'.

Sample	DEL_nr	TEs	SINEs	LINE1s	LINE1_frag	LINE1 >5kb	Other
Polar bear 01	1,324	911	535	251	246	5	125
Polar bear 02	2,005	1,350	723	422	414	7	205
Brown bear	22,976	19,862	14,157	3,004	2,945	59	2,701
Am black bear	23,481	20,877	16,247	2,900	2,829	71	1,730
As black bear	33,458	30,195	22,335	3,954	3,851	103	3,906
Sloth bear	35,029	31,645	23,726	3,998	3,897	101	3,921
Sun bear	33,625	30,563	23,026	3,666	5,568	98	3,871
Spectacled bear 01	72,207	68,181	54,917	6,154	6,000	154	7,110
Spectacled bear 02	71,329	67,105	55,333	6,260	6,105	155	5,512
Total	295,434	270,689	210,999	30,609	29,855	754	29,081

Supplementary Table 12. Heterozygous loci identified by PCR. For each species with >=1 heterozygous PCR amplicons, the number of heterozygous (Het) and homozygous (Hom) loci are indicated. Heterozygosity (% Het) is estimated by dividing the number of heterozygous amplicons by total amplicon count.

Species	Het	Hom	Total	% Het
Polar bear	2	49	51	3.92
Brown bear	8	39	47	17.02
American black bear	3	48	51	5.88
Asiatic black bear	4	67	71	5.63
Sun bear	3	54	57	5.26
Sloth bear	1	55	56	1.79
Total	21	312	333	6.31

Supplementary Table 13. KKSC test results for SINE insertion counts. The KKSC test was performed on two clades, consisting of species triplets. The first clade (PB-BB-AMB) is polar bear, brown bear and American black bear. The most likely tree according to the test is ((Polar bear, Brown bear), American black bear) at p=1.6106E-207. Gene flow is inferred from polar bear and American black bear to brown bear at p=3.2169E-193. The second clade (ASB-SUN-SLO) is Asiatic black, sun and sloth bear and the most likely topology is ((Sun bear, Sloth bear), Asiatic black bear) and hybridization is rejected at p=0.9686.

Clade	Test type	Significance level	Test values	Critical border at p<=0.05	Critical border at p<=0.01
PB-BB-AMB	hybridization test	8.8623E-287	1371 vs 1371 + 99	76	100
	tree test	1.0404E-159	3160 vs 1371	133	174
ASB-SUN-SLO	hybridization test	0.6066	278 vs 278 + 265	47	61
	tree test	6.8845E-47	3993 vs 2809	163	213

Supplementary Data

Supplementary Data 1. Spreadsheet with loci and primer sequences

Supplementary Data 2. Tab-separated file of final TE dataset

Supplementary Data 3. FASTA alignments of selected validated loci (ZIP archive)

Supplementary Notes

Supplementary Note 1- Discrepancies between NGS-generated and Sanger sequences.

The paired-end Illumina sequences of each genome were mapped against the polar bear reference genome, then SNVs were called from the short-read alignments to generate consensus sequences of the resequenced genomes (Kumar et al. 2016). In the consensus sequences, Ref- TE insertions are generally not assembled. The presence of a TSD, flanking the potential TE insertion, was frequently observed in the consensus sequence. Experimental validation using Sanger sequencing of the DNA from the same individual showed that the TSDs flanking the breakpoint were artificially generated during the mapping process and represent artifacts. The presence of TSDs in consensus sequences generated from short read alignments, is therefore not informative regarding the presence of a TE insertions in the resequenced genome. Accordingly, there was also no correlation between the TE prediction calls and the incorrectly generated TSDs in the consensus sequences (**Supplementary Fig. 10**).

Supplementary Note 2. Remarks on flanking sequence analysis.

In addition to test the association between TE insertion loci with flanking substitutions that exhibit the same phylogenetic signal, we screened the flanks for the presence of substitutions that support alternative phylogenetic hypotheses. We extracted flanking sequences shared by:

- S1: Asiatic black bear, Sun bear, and Sloth bear
- S2: Asiatic black bear and American black bear
- S3: Asiatic Black bear & Sloth bear S4: Asiatic Black bear & Sun bear
- S5: Sun bear & Sloth bear

The trees per row (T1-T5) show the selected topologies supported by TE insertions (**Supplementary Fig 13**). In the diagrams, the mean frequency of substitutions supporting phylogenetic signals are shown in 1 kb windows. The columns S1-S5 indicate for which taxa synapomorphic substitutions were selected for the frequency analysis. The blue dots in the trees S1-S5 in which taxa the substitutions are present. As described in the main text, substitutions supporting the phylogeny indicated by TEs accompany a TE insertion loci at different spatial scales (**Supplementary Fig 13 panels in the diagonal**). In general, we observe no substitutions supporting other topologies than indicated by the TE insertion. Interestingly, loci carrying TE insertions in Asiatic black bear and either sun or sloth bear show - in addition to phylogenetically congruent substitutions - elevated substitutions that group together all three Asian bear species (**Supplementary Figure 13 panel T3-S1, T4-S1**). A possible evolutionary model (model 1) to explain this pattern is that species-tree congruent TE loci reflect the speciation history of Asiatic black bear, sun and sloth bear. Introgressive hybridization post-speciation transferred the TE-containing loci between Asiatic black and one of its sister species. A second explanation (model 2) is that the TE insertion is an ancestral polymorphism cannot be ruled out for these cases. However, if the higher amount of substitutions supporting the species-tree phylogeny is considered a consequence of an older common ancestry (i.e. more time has past to accumulate substitutions) rather the first model is supported.

Other deviations between TE insertion and the surrounding substitutions can be observed in panels T1-S5 and T2-S5 (**Supplementary Fig. 13**). In panel T1-S5, in addition to the TE congruent substitutions that are shared by all three Asiatic bear species, an increased number of substitutions supporting the monophyly of sun and sloth bear are found. However, this pattern is expected as it reflects the speciation history of all three species.

In panel T2-S5, loci with TE insertions in American and Asiatic black bear show substitutions supporting also the monophyly of sun and sloth bear. This can be explained by the speciation history where the Asiatic black bear diverged from the ancestor of all three Asian bear species and subsequently exchanged alleles with a lineage related to the American black bear or maintained ancestral polymorphisms from the initial ursine radiation.